Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Estimation sur petits domaines à l’aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d’échantillonnage

par Yong You

Date de diffusion : le 6 janvier 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2022

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 389-399
Statistique Canada, nº 12-001-X au catalogue

Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage

Yong You[1]

Résumé

Nous considérons ici le modèle d'estimation sur petits domaines de Fay-Herriot. Nous nous intéressons en particulier à l'incidence du lissage et de la modélisation de la variance d'échantillonnage sur les estimations par modèle. Nous présentons des méthodes permettant de lisser et de modéliser les variances d'échantillonnage et appliquons les modèles proposés à une analyse de données réelles. Nos résultats font voir qu'un lissage de variance d'échantillonnage est de nature à accroître l'efficacité et la précision de l'estimateur par modèle. Dans une modélisation de variance d'échantillonnage, les modèles hiérarchiques bayésiens de You (2016) et de Sugasawa, Tamae et Kubokawa (2017) améliorent tous aussi bien les estimations d'enquête directes.

Mots-clés : MPLSBE; méthode hiérarchique bayésienne; échantillonnage de Gibbs; modèle loglinéaire; erreur relative; échantillonnage; variance; petit domaine.

1. Introduction

L'estimation sur petits domaines est prisée et importante dans l'analyse de données d'enquête. On a largement recouru aux modèles pour établir des estimations fiables sur petits domaines. Dans la pratique, on se sert d'habitude de modèles au niveau de domaine chaque fois que des estimations d'enquête directes et des variables auxiliaires de niveau de domaine sont disponibles. Divers modèles en ce sens ont été proposés en vue d'accroître la précision des estimations d'enquête directes (voir Rao et Molina, 2015). Parmi les modèles au niveau de domaine, celui de Fay-Herriot (Fay et Herriot, 1979) est un modèle de base amplement utilisé dans l'estimation sur petits domaines. Il comporte deux composantes, à savoir un modèle d'échantillonnage pour les estimations d'enquête directes et un modèle de lien pour le paramètre du petit domaine choisi. Avec le modèle d'échantillonnage, nous posons qu'un estimateur d'enquête direct y_i est sans biais sous le plan de sondage pour le paramètre de petit domaine θ_i, de sorte que

$$y_i = \theta_i + e_i, \quad i = 1, \ldots, m, \tag{1.1}$$

où e_i est l'erreur d'échantillonnage liée à l'estimateur direct y_i et où m est le nombre de petits domaines. L'hypothèse habituelle est que les e_i sont des variables aléatoires normales indépendantes dont la moyenne est $\mathrm{E}(e_i) = 0$ et la variance d'échantillonnage, $\mathrm{Var}(e_i) = \sigma_i^2$. Avec le modèle de lien, nous posons que le paramètre de petit domaine θ_i est lié aux variables auxiliaires $x_i = (x_{i1}, \ldots, x_{ip})'$ par un modèle de régression linéaire qui est

$$\theta_i = x_i' \beta + v_i, \quad i = 1, \ldots, m, \tag{1.2}$$

1. Yong You, Centre de coopération internationale et d'innovation en méthodologie (CCIIM), Statistique Canada, Ottawa, Canada. Courriel : yong.you@statcan.gc.ca.

où $\beta = (\beta_1, \ldots, \beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les v_i sont des effets aléatoires propres au domaine que nous supposons indépendants et d'une distribution identique avec $\mathrm{E}(v_i) = 0$ et $\mathrm{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité est généralement formulée. Les effets aléatoires v_i et les erreurs d'échantillonnage e_i sont mutuellement indépendants. La variance de modèle σ_v^2 est inconnue et doit faire l'objet d'une estimation. La combinaison des modèles (1.1) et (1.2) nous donne un modèle mixte linéaire qui est

$$y_i = x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m. \tag{1.3}$$

Le modèle (1.3) fait intervenir tant des erreurs aléatoires fondées sur le plan e_i que des effets aléatoires fondés sur le modèle v_i. Nous supposons ordinairement que, dans le cas du modèle de Fay-Herriot, la variance d'échantillonnage σ_i^2 est connue. C'est là une hypothèse très forte et, dans la pratique, nous disposerons généralement d'estimations directes sans biais des variances d'échantillonnage. S'il s'agit d'employer des estimations directes de variance d'échantillonnage, deux méthodes s'offrent dans la pratique, soit par lissage et par modélisation. Dans le premier cas, des estimations lissées de ces variances entrent dans le modèle de Fay-Herriot et sont ensuite traitées comme connues. Cette méthode exige des variables et des modèles externes comme une fonction généralisée de variance (FGV) et des effets du plan de sondage. You et Hidiroglou (2012) se sont attachés aux méthodes avec FGV et effets du plan pour un lissage des variances d'échantillonnage en présence de proportions. Nous emploierons ici un modèle FGV proposé par You et Hidiroglou (2012) en matière de lissage de variance d'échantillonnage.

Une solution autre que le lissage est le recours habituel dans la pratique à une modélisation de variance d'échantillonnage. Soit s_i^2 désignant l'estimateur direct de la variance d'échantillonnage σ_i^2. Nous considérons pour s_i^2 un modèle spécial sous la forme $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et où n_i est la taille d'échantillon du i^{e} domaine. Rivest et Vandal (2002) et Wang et Fuller (2003) ont appliqué la méthode dite du meilleur prédicteur linéaire sans biais empirique (MPLSBE) pour dégager les estimations fondées sur le modèle. You et Chapman (2006) ont retenu une méthode hiérarchique bayésienne (HB) et combiné le modèle de variance d'échantillonnage $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ et le modèle de petit domaine (1.3). Le modèle ainsi obtenu tire sa puissance des estimations à la fois de petit domaine et de variance d'échantillonnage. Ainsi, la modélisation intégrée HB avec $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ a largement été utilisée dans la pratique, notamment par You (2008, 2016), Dass, Maiti, Ren et Sinha (2012), Sugasawa, Tamae et Kubokawa (2017), Ghosh, Myung et Moura (2018) et Hidiroglou, Beaumont et Yung (2019).

Nous regarderons ici les approches par lissage et par modélisation des variances d'échantillonnage. À la section 2, nous présentons la méthode MPLSBE en fonction des estimations tant lissées que directes des variances d'échantillonnage. À la section 3, nous présentons le modèle HB de Fay-Herriot et trois autres modèles HB basés sur la modélisation des variances d'échantillonnage. À la section 4, nous comparons les effets du lissage et de la modélisation de la variance d'échantillonnage par une analyse de données réelles et enfin, à la section 5, nous y allons de nos suggestions.

2. Modèle de Fay-Herriot avec le cadre MPLSBE

Si nous employons le modèle de Fay-Herriot (1.3) et posons que σ_i^2 et σ_v^2 sont connus dans ce modèle, nous obtenons le meilleur estimateur de prédiction linéaire sans biais (MPLSB) de θ_i comme $\tilde{\theta}_i = \gamma_i\, y_i + (1-\gamma_i)\, x_i'\tilde{\beta},$ où $\gamma_i = \sigma_v^2/(\sigma_v^2+\sigma_i^2)$ et $\tilde{\beta} = \left(\sum_{i=1}^m (\sigma_i^2+\sigma_v^2)^{-1} x_i\, x_i'\right)^{-1}\left(\sum_{i=1}^m (\sigma_i^2+\sigma_v^2)^{-1} x_i\, y_i\right).$ Pour estimer la composante σ_v^2 de la variance, nous devons supposer au départ que σ_i^2 est connu. Plusieurs méthodes s'offrent pour l'estimation de cette composante; nous allons employer la méthode du maximum de vraisemblance restreint (REML). Nous obtenons ainsi le MPLSBE du paramètre de petit domaine θ_i comme

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\, x_i'\hat{\beta}, \tag{2.1}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2+\sigma_i^2)$ et où $\hat{\sigma}_v^2$ est l'estimateur REML. L'estimateur de l'erreur quadratique moyenne (EQM) de $\hat{\theta}_i$ nous est donné par $\text{eqm}(\hat{\theta}_i) = g_{1i} + g_{2i} + 2g_{3i},$ où $g_{1i} = \hat{\gamma}_i\sigma_i^2$ est le terme principal, où g_{2i} rend compte de la variabilité due à l'estimation du paramètre de régression β et où g_{3i} découle de l'estimation de la variance de modèle σ_v^2 (voir les détails dans Rao et Molina, 2015).

Nous pouvons utiliser l'estimation lissée ou l'estimation directe de σ_i^2 en (2.1). S'il s'agit de lisser la variance d'échantillonnage, nous appliquons un modèle de régression loglinéaire à la variance d'estimation directe $s_i^2,$ comme le proposent You et Hidiroglou (2012). Le modèle de lissage se définit comme

$$\log(s_i^2) = \eta_0 + \eta_1 \log(n_i) + \varepsilon_i, \quad i = 1, \ldots, m, \tag{2.2}$$

où le terme d'erreur du modèle est $\varepsilon_i \sim N(0, \psi^2)$ et où ψ^2 est inconnu. Soit $\hat{\eta}_0$ et $\hat{\eta}_1$ les estimations par les moindres carrés ordinaires des coefficients de régression η_0 et $\eta_1,$ et soit $\hat{\psi}^2$ la variance résiduelle estimée du modèle de régression loglinéaire (2.2). Nous pouvons obtenir un estimateur lissé de la variance d'échantillonnage σ_i^2 sous la forme suivante :

$$\tilde{\sigma}_i^2 = \exp\left(\hat{\eta}_0 + \hat{\eta}_1 \log(n_i)\right)\exp\left(\hat{\psi}^2/2\right).$$

Les variances d'échantillonnage en lissage $\tilde{\sigma}_i^2$ peuvent alors être employées dans l'estimateur MPLSBE (2.1) et le calcul de son EQM. C'est là une procédure courante (voir Rao et Molina, 2015).

Là où une estimation directe de la variance d'échantillonnage s_i^2 remplace sa valeur réelle σ_i^2 en (2.1), un terme supplémentaire rendant compte de l'incertitude de l'utilisation de s_i^2 est nécessaire dans l'estimateur EQM. Ce terme appelé g_{4i} est donné par $g_{4i} = 4(n_i-1)^{-1}\hat{\sigma}_v^4\, s_i^4\, (\hat{\sigma}_v^2 + s_i^2)^{-3}$ (voir Rivest et Vandal (2002) et Rao et Molina (2015), page 150). En employant directement s_i^2 dans le cadre MPLSBE, on risque de surestimer la variance du modèle σ_v^2 (You, 2010; Rubin-Bleuer et You, 2016) et d'obtenir des estimations moins fidèles. Nous comparerons les estimations MPLSBE et HB en fonction des valeurs lissées et directes de variance d'échantillonnage à la section 4.

3. Modèle de Fay-Herriot avec le cadre HB en modélisation de la variance d'échantillonnage

Nous présenterons d'abord le modèle de Fay-Herriot dans un cadre hiérarchique bayésien (HB) et considérerons ensuite trois modèles de la variance d'échantillonnage. Le premier est de You et Chapman (2006) où un modèle à distribution gamma inverse est appliqué à la variance d'échantillonnage σ_i^2 avec des valeurs connues de paramètres vagues. Le deuxième modèle est de You (2016) avec un modèle loglinéaire à erreur aléatoire appliqué à σ_i^2. Le troisième est proposé par Sugasawa et coll. (2017) avec un modèle à distribution gamma inverse appliqué à σ_i^2, mais dans des réglages paramétriques différents.

Modèle HB 1 : modèle de Fay-Herriot dans un cadre HB, que nous appellerons FH-HB :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \;\; \pi(\sigma_v^2) \propto 1.$

À noter que, dans ce modèle, la variance d'échantillonnage σ_i^2 est censée être connue. À la place de σ_i^2, il y aura une estimation lissée de la variance d'échantillonnage $\tilde{\sigma}_i^2$ ou une estimation directe s_i^2.

Modèle HB 2 : modèle de You-Chapman (You et Chapman, 2006), que nous appellerons YCM :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind}\, \sigma_i^2 \chi_{d_i}^2, \;\; d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i),$ où $a_i = 0{,}0001,\; b_i = 0{,}0001,\; i = 1, \ldots, m;$
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \;\; \pi(\sigma_v^2) \propto 1.$

On peut trouver dans You et Chapman (2006) les distributions conditionnelles complètes pour la procédure d'échantillonnage de Gibbs avec les modèles FH-HB et YCM.

Modèle HB 3 : modèle loglinéaire de You (2016) pour les variances d'échantillonnage, que nous appellerons YLLM :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind}\, \sigma_i^2 \chi_{d_i}^2, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\log(\sigma_i^2) \sim N\left(\delta_1 + \delta_2 \log(n_i), \tau^2\right), \quad i = 1, \ldots, m;$

- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \quad \pi(\delta_1, \delta_2) \propto 1, \quad \pi(\sigma_v^2) \propto 1, \quad \pi(\tau^2) \propto 1.$

Signalons que le modèle YLLM est un modèle loglinéaire appliqué à la variance d'échantillonnage σ_i^2 qui amplifie le modèle proposé par Souza, Moura et Migon (2009) en utilisant $\log(n_i)$ et en ajoutant un effet aléatoire à la partie « régression » du modèle. On trouvera en annexe les distributions conditionnelles complètes pour la procédure d'échantillonnage de Gibbs.

Modèle HB 4 : modèle de Sugasawa, Tamae et Kubokawa (2017) resserrant tant les moyennes que les variances, que nous appellerons STKM :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,\big|\, \sigma_i^2 \sim \text{ind}\, \sigma_i^2 \chi_{d_i}^2, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i\gamma),$ où a_i et b_i sont des constantes connues, $a_i = O(1); \; b_i = O(n_i^{-1});$
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \quad \pi(\sigma_v^2) \propto 1, \quad \pi(\gamma) \propto 1.$

Mentionnons que, dans STKM pour le modèle à distribution gamma inverse de σ_i^2, nous choisissons $a_i = 2$ et $b_i = n_i^{-1}$, ainsi que le proposent Sugasawa et coll. (2017). Ghosh et coll. (2018) ont opté pour le même paramétrage dans leur étude comparative des estimateurs HB. On peut trouver les distributions conditionnelles complètes pour STKM dans Sugasawa et coll. (2017).

À noter que, dans les modèles HB 2 à 4 qui précèdent, la modélisation khi-carré de la variance d'échantillonnage $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ fait appel à la normalité et à un échantillonnage aléatoire simple (Rivest et Vandal, 2002). Dans le cas des plans de sondage complexes, on pourrait devoir déterminer plus soigneusement les degrés de liberté d_i. Il n'y a pas de solide résultat théorique pour cette détermination (Dass et coll., 2012). Des formules peuvent se révéler utiles comme l'approximation sur erreurs de niveau d'unité non normales que proposent Wang et Fuller (2003) ou la règle par simulation de Maples, Bell et Huang (2009), mais on aura besoin dans l'un et l'autre cas de données de niveau des unités et d'une vaste étude de simulation. Une détermination prudente des degrés de liberté pourrait constituer une approximation raisonnablement utile. Ajoutons qu'une analyse bayésienne d'ajustement de modèle pourrait aussi servir à une détermination du modèle.

4. Application

Dans cette section, nous appliquons les modèles des sections 2 et 3 aux données canadiennes de l'Enquête sur la population active (EPA) et comparons les estimations MPLSBE et HB. L'EPA diffuse

chaque mois des estimations du taux de chômage pour de grands territoires comme le pays entier et les provinces aussi bien que pour de petites régions comme les régions métropolitaines de recensement (RMR) et les agglomérations de recensement (AR) de tout le Canada. Les estimations directes de l'EPA pour un certain nombre d'unités locales manquent de fiabilité, présentant des coefficients de variation (CV) très élevés à cause des petites tailles d'échantillon. Nous considérons que les estimateurs fondés sur le modèle viennent améliorer ces estimations directes. En guise d'illustration, nous appliquons le modèle de Fay-Herriot aux estimations du taux de chômage de mai 2016 au niveau des RMR-AR; nous comparons les estimations fondées sur le modèle et directes aux estimations du recensement pour mettre en contraste les effets du lissage et de la modélisation de la variance d'échantillonnage. Hidiroglou et coll. (2019) ont aussi comparé les estimations de l'EPA fondées sur le modèle aux estimations du recensement. Dans l'estimation du taux de chômage, la proportion mensuelle de prestataires locaux de l'assurance-emploi sert de variable auxiliaire dans le modèle. Dans une comparaison des estimations ponctuelles, nous calculons l'erreur relative absolue (ERA) des estimations directes et fondées sur le modèle relativement aux estimations du recensement pour chaque RMR ou AR :

$$\text{ERA}_i = \left| \frac{\theta_i^{\text{Census}} - \theta_i^{\text{Est}}}{\theta_i^{\text{Census}}} \right|,$$

où θ_i^{Est} est l'estimation directe ou par MPLSBE/HB et où θ_i^{Census} est la valeur correspondante du recensement pour le taux de chômage. Il s'agit ensuite de prendre la moyenne des ERA sur l'ensemble des RMR et AR. Nous calculons les CV moyens des estimations directes et fondées sur le modèle. Nous privilégierons un modèle où les ERA et les CV sont moindres.

Nous appliquons les modèles d'abord aux 117 RMR et AR à taille d'échantillon ≥ 2, ensuite à 92 à taille d'échantillon ≥ 5 et enfin à 79 à taille d'échantillon ≥ 7. Le tableau 4.1 présente les ERA moyennes et les CV moyens correspondants (ceux-ci entre parenthèses). Dans ce tableau, les valeurs présentées de modélisation sont fonction du lissage ou de l'estimation directe de la variance d'échantillonnage.

En cas de lissage, les deux méthodes FH-MPLSBE et FH-HB améliorent nettement les estimations d'enquête directes avec les ERA et les CV sont bien moindres. Mentionnons en particulier que l'ERA et le CV sont respectivement les moindres avec FH-HB et FH-MPLSBE. Si la modélisation porte sur les 117 régions, les ERA et CV moyens sont respectivement de 0,263 et 0,329 pour l'estimateur direct de l'EPA, de 0,124 et 0,087 pour le lissage FH-MPLSBE et de 0,118 et 0,116 pour le lissage FH-HB. Le bon résultat des lissages FH-MPLSBE et FH-HB de la variance d'échantillonnage indique que la fonction généralisée de variance (2.2) est très utile dans ce cas et réussit à améliorer les estimations fondées sur le modèle.

En cas d'estimation directe de la variance d'échantillonnage, FH-MPLSBE et FH-HB ont les pires résultats de tous les modèles, et ils sont presque identiques dans ce scénario. Les trois autres modèles HB s'en tirent mieux que les modèles FH-MPLSBE et FH-HB en cas d'estimation directe. YLLM et STKM sont d'un meilleur rendement que YCM avec des ERA et des CV moindres. YLLM et STKM ont des

résultats à peu près identiques pour tous les groupes de RMR-AR; YLLM montre toujours une ERA légèrement inférieure, mais un CV légèrement supérieur à ceux de STKM. Sur l'ensemble des 117 régions, YLLM et STKM ont, par exemple, des ERA à 0,135 et 0,137 et des CV moyens à 0,123 et 0,122. YCM présente des valeurs correspondantes de 0,148 et 0,136 et FH-HB, de 0,171 et 0,221.

Tableau 4.1
Comparaison des erreurs relatives absolues (ERA) et des coefficients de variation (CV) en moyenne avec les valeurs CV entre parenthèses

RMR-AR	Estimation directe EPA	FH-MPLSBE lissage	FH-HB lissage	FH-MPLSBE estimation directe	FH-HB estimation directe	YCM estimation directe	YLLM estimation directe	STKM estimation directe
Moyenne sur les 117 RMR-AR	0,263	0,124	0,118	0,170	0,171	0,148	0,135	0,137
(taille d'échantillon ≥ 2)	(0,329)	(0,087)	(0,116)	(0,238)	(0,221)	(0,136)	(0,123)	(0,122)
Moyenne sur 92 RMR-AR	0,216	0,124	0,116	0,133	0,132	0,132	0,125	0,127
(taille d'échantillon ≥ 5)	(0,262)	(0,076)	(0,103)	(0,123)	(0,123)	(0,121)	(0,117)	(0,116)
Moyenne sur 79 RMR-AR	0,181	0,122	0,113	0,126	0,122	0,122	0,118	0,120
(taille d'échantillon ≥ 7)	(0,232)	(0,057)	(0,094)	(0,115)	(0,115)	(0,115)	(0,114)	(0,113)

Passons maintenant à une comparaison de modélisation bayésienne avec ordonnée prédictive conditionnelle (OPC) qui porte sur les quatre modèles HB pour l'estimation directe. Les OPC sont les valeurs observées de vraisemblance en fonction de la distribution prédictive sur validation croisée $f(y_i | y_{\text{obs}(i)})$. Nous calculons les valeurs OPC pour les divers points $y_{i,\text{obs}}$ de données d'observation. Une OPC plus élevée indique que $y_{i,\text{obs}}$ favorise le modèle et permet un meilleur ajustement. S'il s'agit de choisir un modèle, nous pouvons calculer le rapport des OPC entre un modèle A et un modèle B. Si ce rapport est supérieur à 1, $y_{i,\text{obs}}$ favorise le modèle A. Nous calculons les rapports OPC pour YCM/FH-HB, YLLM/FH-HB et STKM/FH-HB et comptons les fois que le rapport est supérieur à 1. Nous pouvons aussi tracer la courbe des valeurs OPC ou condenser celles-ci en prenant la moyenne des OPC estimées. Pour plus de détails sur l'ordonnée prédictive conditionnelle, voir, par exemple, Gilks, Richardson et Spiegelhalter (1996), page 153, You et Rao (2000) et Molina, Nandram et Rao (2014). Le tableau 4.2 présente les OPC moyennes et médianes pour les 117 RMR-AR, ainsi que le nombre de rapports OPC supérieurs à 1.

Tableau 4.2
Résumé des valeurs et des rapports OPC pour les 117 RMR-AR

	FH-HB estimation directe	YCM estimation directe	YLLM estimation directe	STKM estimation directe
OPC moyenne	0,1053	0,1222	0,1242	0,1238
OPC médiane	0,0976	0,1004	0,1045	0,1051
nombre de rapports OPC supérieurs à 1	-	72	78	76

Il ressort du tableau 4.2 que YCM, YLLM et STKM présentent des valeurs OPC supérieures à celles de FH-HB, indice que le modèle HB en modélisation de variance d'échantillonnage est préférable lorsque les estimations directes de cette variance sont employées. Il ressort également que YLLM et STKM sont préférables à YCM. Dans le cas des rapports OPC sur l'ensemble des 117 régions, 72 régions ou observations favorisent YCM, 78 YLLM et 76 STKM. Ainsi, plus d'observations font préférer YCM, YLLM et STKM à FH-HB; YLLM présente le plus de rapports OPC supérieurs à 1. La comparaison OPC s'accorde avec les résultats au tableau 4.1. Pour les autres méthodes de vérification et d'évaluation de modèles, voir Hidiroglou et coll. (2019).

5. Conclusion

Nous avons comparé les estimations fondées sur le modèle de Fay-Herriot dans les cas respectifs de lissage et de modélisation des variances d'échantillonnage. Comme dans Hidiroglou et coll. (2019), nos résultats indiquent que ce modèle peut grandement améliorer les estimations directes du taux de chômage de l'EPA, bien que des modèles plus complexes tels que les modèles non appariés ou les modèles de séries chronologiques puissent être employés (You (2008), par exemple). De tous les estimateurs, ce sont les modèles FH-MPLSBE et FH-HB en lissage qui donnent le meilleur résultat de réduction des ERA et des CV. En estimation directe des variances d'échantillonnage, ces mêmes modèles donnent les pires résultats. Dans la modélisation HB, YLLM et STKM s'en tirent tous deux très bien et mieux que YCM; YLLM est un peu meilleur que STKM dans l'étude. Nous proposons le modèle YLLM ou STKM en cas d'estimation directe de la variance d'échantillonnage. Autre possibilité, le lissage de variance devrait permettre dans le modèle de Fay-Herriot de surmonter la difficulté de modélisation des variances d'échantillonnage que nous évoquons à la section 3. Le lissage avec le modèle FGV en (2.2) à la section 2 peut produire un très bon résultat, comme le démontre notre étude.

Annexe

Distributions conditionnelles complètes et procédure d'échantillonnage pour YLLM

- $\left[\theta_i \,|\, y, \beta, \sigma_i^2, \sigma_v^2\right] \sim N\left(\gamma_i y_i + (1-\gamma_i)\, x_i'\beta, \gamma_i \sigma_i^2\right)$, où $\gamma_i = \sigma_v^2 / \sigma_v^2 + \sigma_i^2$, $i = 1, \ldots, m$;
- $\left[\ \beta \,|\, y, \theta, \sigma_i^2, \sigma_v^2\ \right] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i \theta_i\right), \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right)$;
- $\left[\ \sigma_v^2 \,|\, y, \theta, \beta, \sigma_i^2\ \right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2}\sum_{i=1}^m (\theta_i - x_i'\beta)^2\right)$;
- $\left[\ \sigma_i^2 \,|\, y, \theta, \beta, \sigma_v^2, \delta, \tau^2\ \right] \propto f(\sigma_i^2) \cdot h(\sigma_i^2)$, où $f(\sigma_i^2)$ et $h(\sigma_i^2)$ sont $f(\sigma_i^2) \sim \mathrm{IG}\left(\frac{d_i + 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right)$ et $h(\sigma_i^2) = \exp\left(-\frac{(\log(\sigma_i^2) - z_i'\delta)^2}{2\tau^2}\right)$;
- $\left[\ \delta \,|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \tau^2\ \right] \sim N_2\left(\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\left(\sum_{i=1}^m z_i \log(\sigma_i^2)\right), \tau^2\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\right)$;

- $\left[\tau^2 \,\middle|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \delta \right] \sim \text{IG}\left(\frac{m}{2} - 1, \frac{1}{2} \sum_{i=1}^{m} \left(\log(\sigma_i^2) - z_i' \delta \right)^2 \right)$.

Nous passons par l'étape de rejet de l'algorithme Metropolis-Hastings pour actualiser σ_i^2 :

(1) on tire σ_i^{2*} de $\text{IG}\left(\frac{d_i + 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \right)$;

(2) on calcule la probabilité d'acceptation $\alpha(\sigma_i^{2*}, \sigma_i^{2(k)}) = \min\left\{ h(\sigma_i^{2*}) / h(\sigma_i^{2(k)}), 1 \right\}$;

(3) on tire u de la distribution uniforme $(0, 1)$; si $u < \alpha(\sigma_i^{2*}, \sigma_i^{2(k)})$, la valeur candidate σ_i^{2*} est acceptée et $\sigma_i^{2(k+1)} = \sigma_i^{2*}$; dans le cas contraire, σ_i^{2*} est rejeté et $\sigma_i^{2(k+1)} = \sigma_i^{2(k)}$.

Remerciements

J'aimerais remercier le rédacteur en chef, le rédacteur adjoint et un examinateur de leurs observations et leurs suggestions constructives ayant permis d'améliorer cet article.

Bibliographie

Dass, S.C., Maiti, T., Ren, H. et Sinha, S. (2012). Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances. *Techniques d'enquête*, 38, 2, 187-203. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11756-fra.pdf.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., Myung, J. et Moura, F.A.S. (2018). Estimation bayésienne robuste sur petits domaines. *Techniques d'enquête*, 44, 1, 109-124. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54959-fra.pdf.

Gilks, W.R., Richardson, S. et Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Maples, J., Bell, W. et Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.

Molina, I, Nandram, B. et Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics*, 8, 852-885.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, 2nd Edition. New York: John Wiley & Sons, Inc.

Rivest, L.P., et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 10 au 13 juillet 2002, Ottawa, Canada.

Rubin-Bleuer, S., et You, Y. (2016). Comparaison de certains estimateurs de variance positifs pour le modèle d'estimation sur petits domaines Fay-Herriot. *Techniques d'enquête*, 42, 1, 69-93. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2016001/article/14542-fra.pdf.

Souza, D.F., Moura, F.A.S. et Migon, H.S. (2009). Prédiction de la population de petits domaines au moyen de modèles hiérarchiques. *Techniques d'enquête*, 35, 2, 203-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009002/article/11042-fra.pdf.

Sugasawa, S., Tamae, H. et Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 1, 21-31. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10614-fra.pdf.

You, Y. (2010). Small area estimation under the Fay-Herriot model using different model variance estimation methods and different input sampling variances. Document de travail de la Direction de la méthodologie, SRID-2010-003E, Statistique Canada, Ottawa, Canada.

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Document de travail de la Direction de la méthodologie, ICCSMD-2016-03-E, Statistique Canada, Ottawa, Canada.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 1, 107-114. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf.

You, Y., and Hidiroglou, M. (2012). Sampling variance smoothing methods for small area proportion estimators. Document de travail de la Direction de la méthodologie, SRID-2012-08E, Statistique Canada, Ottawa, Canada; Présenté en tant que conférencier invité au Fields Institute Symposium on the Analysis of Survey Data and Small Area Estimation, Carleton University, Ottawa, 2012.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 2, 197-206. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000002/article/5537-fra.pdf.